4.3 A 32nm Westmere-EX Xeon® Enterprise Processor

Shankar Sawant, Utpal Desai, Gururaj Shamanna, Lokesh Sharma, Mandar Ranade, Anil Agarwal, Sampath Dakshinamurthy, Rajagopal Narayanan

Intel, Bangalore, India

The next-generation enterprise Xeon® processor [1] consists of 10 Westmere 32nm cores [2] and a shared inclusive L3 cache (LLC) integrated on a monolithic die, with link-based I/Os. This paper focuses on the innovations and circuit optimizations over the predecessor [3] targeting idle power reduction, robust high-speed I/O links, and performance per watt improvements. The processor is implemented in 32nm CMOS using high-κ metal gate transistors and nine copper interconnect layers [4].

Core C6 power management is supported with individually controlled power gates for each of the 10 cores (Fig. 4.3.1). Core state is retained in a dedicated SRAM block, while the power gates are switched-off. Macro clock gating (MCG) is implemented in the uncore to reduce switching power from local clock network and associated logic. Power consumption in this state is limited to critical uncore circuits that monitor incoming snoops, interrupts, and machine check events. The on-die power controller (PCU) controls entry/exit from this mode. The MCG mode signal is overloaded on existing fine-grained DFX infrastructure (Fig 4.3.2) to maximize circuit shut-down granularity right down to the regional clock buffers, synchronized to a common clock edge. An additional level of idle power reduction is provided by extending the package C6 states to disable memory I/O links (SMI-Disable) by shutting off the bias currents to transmit-driver, receiver-VOC, phase-interpolators and DLL. The PC6 state (the deepest processor sleep state) entry is coordinated by CPU firmware to negotiate a guaranteed platform idle. SMI-disable is a sub-state of PC6 leveraging memory link error, fast reset init sequence to exit from this state. To minimize exit latency, several components of the full reset sequence like the VOC calibrate, and voltage offset tuning are bypassed in this exit mode. Also the IO-PLL and compensation loops are kept alive to shorten link bring-up, by saving on PLL lock times. Green VID (GVID) flow is implemented in High Volume Manufacturing (HVM) to minimize uncore operating voltage to further optimize idle power. Unlike the conventional approach to bin parts at highest voltage meeting TDP power, the GVID flow bins parts with highest voltage meeting idle power and TDP power.

The uncore is designed to reduce dynamic power usage, to maximize available TDP budget for ten cores. Fine grain clock gating is done in datapath structures and automated power aware algorithms are used as part of the control block synthesis implementation. Contrary to conventional wisdom a low-leakage process is NOT selected for this multi-core processor. Power distribution numbers (Fig. 4.3.3) show that the dynamic power component of the total TDP is significant, indicating the need to operate uncore at a lower voltage. A low leakage process would reduce intrinsic transistor performance, which is compensated by increasing the voltage to maintain frequency parity, and hence increase overall processor power.

To enable robust I/O data rate over 20" FR4 channel, 2nd order CTLE with temperature compensation and 3-stage clock amplifier with inbuilt duty-cycle corrector (DCC) circuits have been implemented in link's receiver with 1100mV_{pp} differential TX driver swing. The 2nd order CTLE is implemented as two-stage amplifier (Fig. 4.3.4). A 1st stage amplifier is implemented with controllable source-degeneration resistance (R1) and capacitor(C1). The 1st stage provides required peaking, or delta between low frequency and high frequency gain, by reducing DC/low frequency gain. A 2nd stage amplifier is implemented with controllable resistance (R2) across the diode-connected load. The 2nd stage provides higher gain and bandwidth at the operating frequencies of interest, via its peaking response, while maintaining the DC gain similar to a wide-band amplifier. The combination of 1st and 2nd stage DC gain and peaking controllability achieve desired performance for given interconnect and to solve saturation issues for shorter channels. Several per-lane features are implemented in the I/O ports to optimize performance/power and to improve yield, like; per-lane termination resistance control, TX/RX EQ coefficient, bias current controls and DCC. The I/O links speed is decoupled from the uncore voltage by retaining only 20% of the I/O digital control circuits on the high-speed I/O clock. A timeslot-valid generated based on the Bresenham Algorithm is used to qualify the nominal uncore clock to derive a gated uncore clock (GUCLK), that runs at half the I/O-link frequency, for the bulk of the I/O port (80%). Iso-performance is achieved by doubling the datapath width to compensate for the slower uncore clock. The decoupling enables scaling the uncore voltage to achieve low power SKUs while maintaining the full I/O link rate. Statistical DOE models indicate link performance is impacted due to jitter amplification in the off-chip link. The I/O clock design focused on reducing the forwarded (Tx) clock jitter. A per-lane DCC feature is added to recover duty-cycle distortion in the Tx clock (Fig 4.3.5a). Post-silicon results show an improvement of ~2% in the duty cycle of the I/O clocks with this feature. The Tx clock is implemented as pseudo-differential while the Rx clock is low-swing fully differential. Other jitter reduction techniques include jitter-attenuating DLLs on Rx clock path, and improving the bias voltages for Tx buffers. Figure 4.3.5b shows the forwarded I/O clock, captured by a BERT, on an electrical validation system that has a 7" trace. The clock eye is 106ps/420mV at 10^{-6} BER and 87.5ps/406mV at 10^{-15} BER.

The clock generation leverages the implementation from the 8-core predecessor and extends it for the two additional cores. Once powered on, the frequency of the various clock domains is fixed, except for the cores, enabling on-demand performance increase through the turbo-boost™ feature. The output clock from the uncore PLL is first distributed vertically using a binary tree distribution. Matched distributions, embedded in horizontal spines, with their point of divergence at the vertical spine, further distribute the global UCLK. This topology reduces the POD between clock end points and eliminates the need to use deskew compensators used in prior designs, providing an iso-skew profile at 10% lower global clock distribution power.

The processor memory system supports 3 levels of on-die cache hierarchy. L3 cache is organized as 10 independent SRAM modules, each 24 way set-associative, connected via ring bus [2] architecture. 0.212μm² 6T bit-cell is used in L3 cache data array to achieve maximum array efficiency and to minimize SRAM leakage power. 0.272μm² 6T bit-cell is used in L3 cache tag/status arrays to support back-back read/write operation. L3 cache achieves a standby VMIN of 700mV and an active VMIN of 800mV. Inline DECTED in L3 data and SECDED in L3 tag, error correction schemes are incorporated in arrays to prevent data corruption due to soft error and erratic bit failures. Segmented sleep techniques [6] implemented in L3 cache provide leakage power savings of 1.7W without degrading processor frequency. Core and cache recovery schemes are supported to realize lower configuration parts.

Power/iiming convergence of ring bus paths [5] was a prime concern due to ring bus width (~1000 signals) and 2× higher interconnect delay in 32nm over 45nm node. To achieve a balance between timing/power vectors, a dual Vt de-coupled master-slave flop is designed. This dual Vt MS flip-flop achieves a sub-10ps Tsetup time for both rise/fall transitions by controlled skewing of clock inverters I3A and I3B (Fig. 4.3.6) and also achieves power savings due to 60% reduction in data pin input capacitance. Inverter I5 in the de-coupled dual Vt MS flip flop helps eliminate data write-back problem existing in conventional MS flops when master latch opens on rising edge of clock. This 10 core processor with its targeted idle power reduction techniques, and emphasis on robust I/O links and performance per watt, provides a refresh to its predecessor 8 core design.

Acknowledgements:
The authors gratefully acknowledge the work of the talented and dedicated Intel team that implemented this processor.

References:

[1] D Nagaraj, et al., "Westmere-Ex: A 20 thread server CPU", Hot Chips 2010.
[2] N Kurd, et al, "Westmere: A family of 32nm IA processors", *ISSCC Dig. Tech. Papers*, Feb. 2010.
[3] S Rusu, et al., "A 45nm 8-Core Enterprise Xeon Processor", *ISSCC Dig. Tech. Papers*, Feb. 2009.
[4] S. Natarajan, et al., "A 32nm Logic Technology Featuring 2nd-Generation High-k + Metal-Gate Transistors, Enhanced Channel Strain and 0.171um² SRAM Cell Size in a 291Mb Array", *IEDM Tech. Digest*, Dec. 2008.
[5] C Park, et al., "A 1.2TB/s On-Chip Ring Interconnect for 45nm 8-Core Processor", *ISSCC Dig. Tech. Papers*, Feb. 2010.
[6] Y. Wang, et al., "A 4.0 GHz 291 Mb Voltage-Scalable SRAM Design in a 32 nm High-k + Metal-Gate CMOS Technology", *IEEE J. Solid-State Circuits*, vol 45, Issue1, Jan. 2010.

 978-1-61284-302-5/11/$26.00

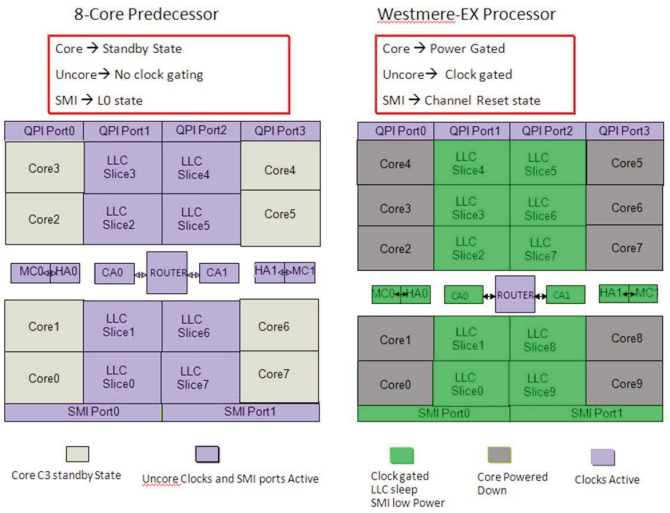


Figure 4.3.1: Idle power reduction features in block diagram.

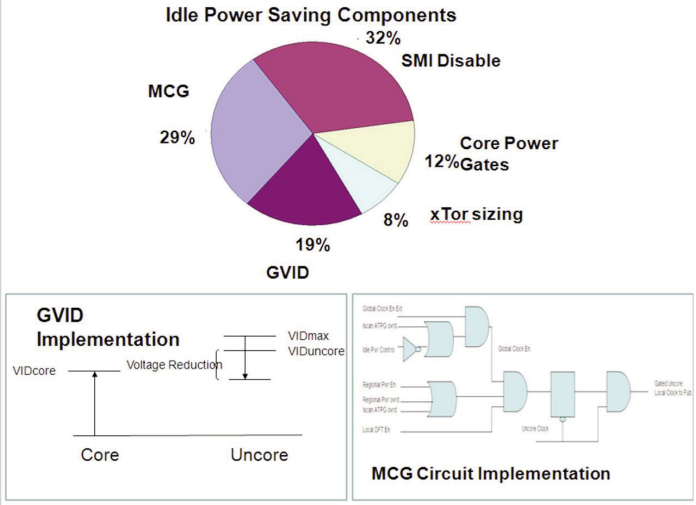


Figure 4.3.2: Idle power reduction techniques, and their contribution to overall idle power savings.

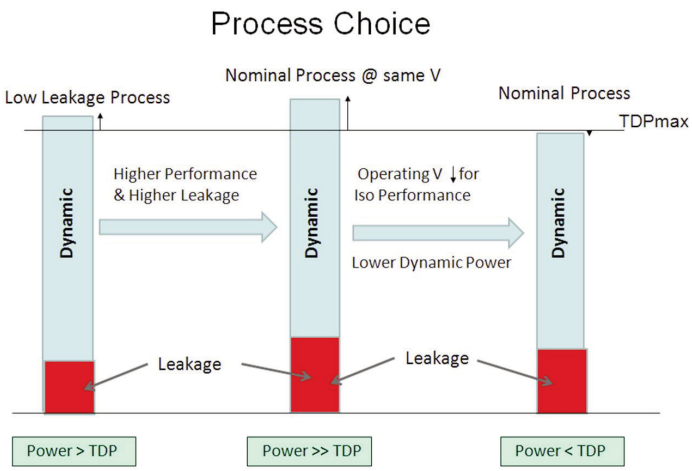


Figure 4.3.3: Process choice for this multi-core processor.

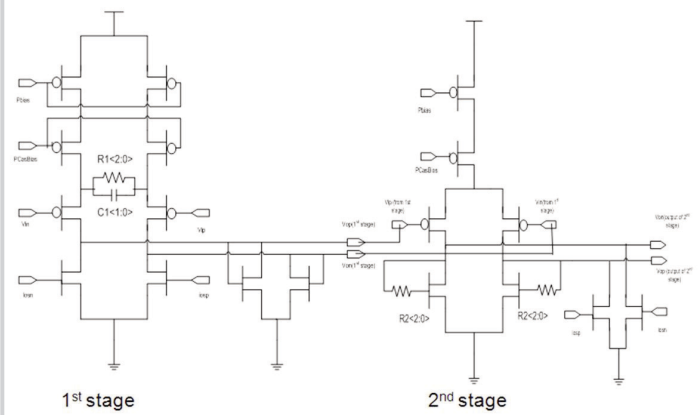


Figure 4.3.4: 2nd order CTLE circuit, with follow-on 2nd stage to increase gain of the amplifier.

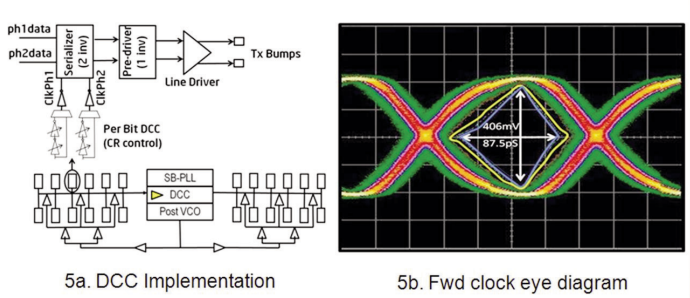
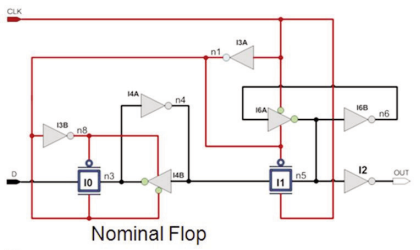


Figure 4.3.5: Duty cycle correction circuit implementation, and measured forwarded clock eye diagram.



Nominal	P-Size	N-Size	TYPE
I0	1.70	1.70	High Vt
I1	2.58	2.58	High Vt
I2	10.73	9.76	High Vt
I3A	1.16	1.30	High Vt
I3B	0.36	0.24	High Vt
I4A	4.40	4.02	High Vt
I4B	1.13	0.78	High Vt
I6A	0.60	0.40	High Vt
I6B	0.14	0.12	High Vt

Modified	P-Size	N-Size	TYPE
I0	0.80	1.00	Low Vt
I1	1.60	1.60	Low Vt
I2	9.00	9.00	High Vt
I3A	0.30	0.45	High Vt
I3B	0.16	0.14	High Vt
I3C	0.16	0.14	High Vt
I4A	0.16	0.14	Low Vt
I4B	0.24	0.18	Low Vt
I5	2.00	2.00	Low Vt
I6A	0.24	0.18	Low Vt
I6B	0.16	0.14	Low Vt

Figure 4.3.6: Modified flop used in the ring bus design.